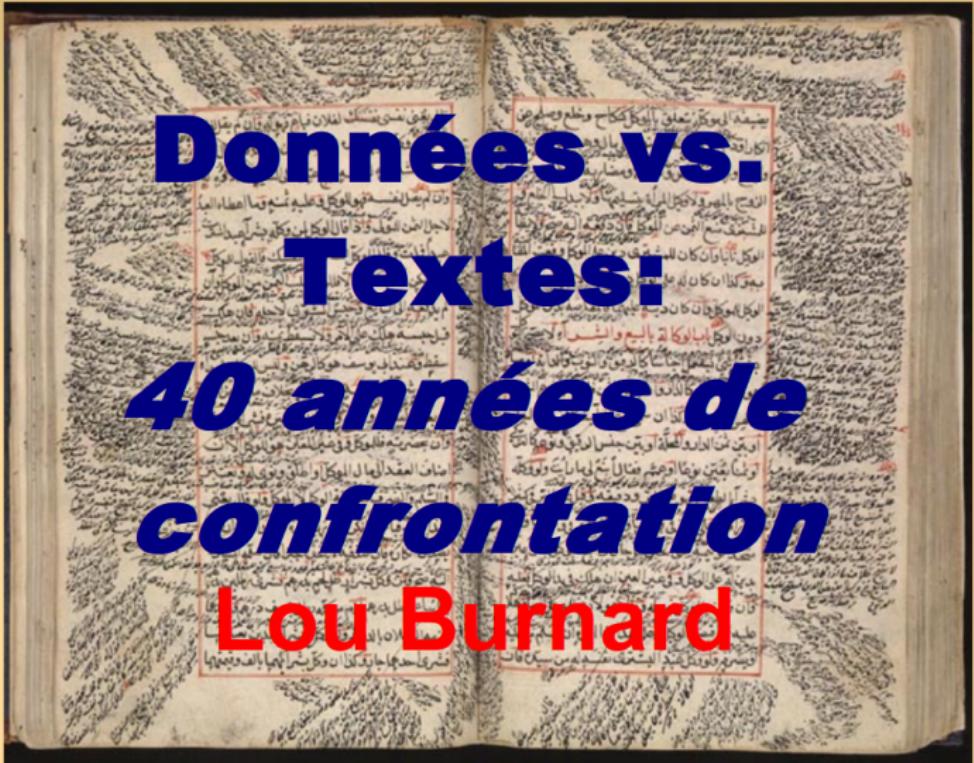


# De 'Literary and Linguistic Computing' jusqu'aux 'Humanites numeriques' quelle importance pour la science des langues?

Lou Burnard



# Plan

- Texte vs Données
- Les trois ages
  - *Literary and Linguistic Computing*
  - *Humanities Computing*
  - *Digital Humanities*
- Quel est ce bruit dans la bibliothèque numérique?



## Textes numériques vs. données numériques

- Le traitement informatisé des données concerne les chiffres, les quantités, les tendances statistiques...
- Le traitement informatisé des textes concerne les mots, l'écriture, la langue...
- L'informatique a donc systématiquement opposé les "donnees" aux "textes"
- en traitant les textes comme si elles étaient des données
- en traitant les donnees comme si elles n'étaient pas des textes

(cf Burnard, 1984)

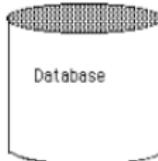
# Three kinds of software



- \* word processing
- \* indexing
- \* database



aardvark
abacus
abracadabra
acidulation
adumbrate
aetherial
affection
agony
aha
ai
ajacanthus



# eBooks

```
div rend="linenumber" xml:id="A4.1">
<head>Beowulf</head>
<bibl>Dobbie, 1953 3-98; Dobbie, E.V.K., Beowulf and Judith, A
<l>Hwæt! We Gardena <caesura/> in geardagum, </l>
<l>þeodcyninga, <caesura/> þrym gefrunon, </l>
<l>hu ða æpelingas <caesura/> ellen fremedon. </l>
<l>Oft Scyld Scfing <caesura/>
    <unclear>seacpene</unclear> breamum, </l>
<l>monegum maegnum, <caesura/> meodosetla ofteah, </l>
<l>egsode <unclear>eorlas</unclear> <caesura/> Syððan arest w
<l>feasceafit funden, <caesura/> he þas frofre gebad, </l>
<l>weox under wolcnum, <caesura/> weorðmyndum þah, </l>
<l>oþþat him aghwylc <caesura/> para ymbisittendra </l>
<l>ofer hronrade <caesura/> hyran scoldé, </l>
<l>gomban gyldan. <caesura/> þat was god cyning! </l>
<l>ðæm eafera was <caesura/> after cenned, </l>
<l>geong in geardum, <caesura/> þone god sende </l>
<l>folce to frofre; <caesura/> fyrendearfe ongaet </l>
<l>þe hie ar drugin <caesura/>
<unclear>aldorleas</unclear>
```

The screenshot shows a digital edition of Beowulf. At the top, there are navigation buttons for 'Library', 'Search', and 'Text View'. The title 'Anglo-Saxon poetic records [Electronic resource]' is displayed. Below the title, the section '137. Beowulf' is shown. The text is presented in two columns. The left column contains the original Old English text with some words highlighted in blue, and the right column contains a modern English translation. The text is numbered in the left margin at 5 and 10. The TEI logo is visible in the bottom left corner.

137. Beowulf

Dobbie, 1953 3-98; Dobbie, E.V.K., *Beowulf and Judith*, ASPR 4 (New York).

Hwæt! We Gardena in geardagum,  
þeodcyninga, þrym gefrunon,  
hu ða æpelingas ellen fremedon.  
Oft Scyld Scfing **seacpene** breamum,  
5 monegum maegnum, meodosetla ofteah,  
egsode **eorlas**. Syððan arest wearð  
feasceafit funden, he þas frofre gebad,  
weox under wolcnum, weorðmyndum þah,  
oþþat him aghwylc 10 para ymbisittendra  
ofer hronrade hyran scoldé,  
gomban gyldan. þat was god cyning!  
ðæm eafera was æfter cenned,

- Les textes numériques se présentent comme des livres imprimés... mais on ne doit pas se laisser séduire par les métaphores !
- Est-ce qu'on numérise les textes juste pour le plaisir de les distribuer dans un nouveau marché avec une nouvelle technologie ? .

# Conclusions

- Il n'y a pas de retour sur le tournant numérique: les infrastructures scientifiques sont désormais numérisées
- Les modèles économiques de l'infrastructure scientifique sont en train d'évoluer
- Les changements plutôt quantitatifs apportés par le numérique provoquent des changements qualitatifs.
- La numérisation massive rend possible de nouveaux perspectives sur la langue.

# Le numérique incontournable

- Les objets de recherches dans les SHS sont devenus numériques
- Les méthodes des SHS ne peuvent pas échapper à l'environnement technologique qui nous englobe
- Une transition du “web de documents” vers le “web de données” s'effectue actuellement
- Les questions politiques et culturelles restent, mais leur contexte évolue dans un monde de plus en plus “ouvert”
- Pour commencer, une petite leçon d'histoire...

# *Literary & Linguistic Computing*



# 1949-1980

- L'age des héros ...
  - Padre Busa et l'Index Thomisticum
  - The Brown Corpus
  - Thesaurus Linguae Graecae
  - etc.
- Concordances, analyse stylistique, études sur l'auctorialité, corpus de langue
- L'ordinateur central géré par des ingénieurs sérieux en blouse blanche lançait des travaux en batch qui étaient transmis à une file d'attente puis exécutés pour produire des sorties

# Colloque ALLC No. 6 (1980)

As an indication of the range of subjects and representatives at the conference, I shall just mention some papers which interested me. Among impressive new textual projects reported on were the collation of the six editions of Burton's Anatomy of Melancholy (Faulkner, Washington State); the indexing of two early 17th century German newspapers by topic (Ries, Cambridge); the production of a lemmatised concordance to Ibsen (Hofland, Bergen) and the problems of concording the textually complex 1606 folio of Ben Jonson (Howard-Hill, South Carolina). There was of course a paper on OCP (Hockey, Oxford) and another on a very quick and very dirty indexing program called CODOC (Niblett, Swansea). There was little else on software of note though there was much informal praise for SPIRES, UNIX and other unattainable goodies. The most impressive hardware on view was that attached to the Chinese Languages Transposition Project (Nancarrow, Cambridge) which hooks a tektronix up to a rotating cylinder for transput of any of several thousand characters in Chinese, Tibetan etc. Among more technical papers, Cercone (British Columbia) surveyed current methods of storing lexicons for natural language applications and Skolnik (Amsterdam) gave a good account of storage mechanism well suited to them. Another new statistical measure of lexical diversity was proposed by Delcourt Mathonet & Mersh (Liege) and the poetic style of W.B.Yeats resisted all attempts to analyse its variations with EYEBALL (Jaynes, Minnesota). Not so Dostoevski, who has now come under the searching eyes of Geir Kjetsaa (Oslo) and his attribution algorithms.

# Problèmes d'auctorialité

*The Occurrence of Common Words in Sentences in THE PAULINE EPISTLES*

No. of Words in Sentence	Rom.	I Cor.	II Cor.	Gal.	Eph.	Phil.	Col.	I Thess.	I Tim.	II Tim.	Heb.
<i>"En"</i>											
No "En"	449	504	232	147	46	56	33	46	78	61	263
One	104	90	71	26	25	30	24	23	20	21	41
Two	18	26	21	8	11	11	19	4	7	5	9
Three	9	6	3	—	9	3	3	7	2	2	2
Four	1	1	2	—	4	1	1	—	—	—	—
Five	—	1	1	—	2	—	1	—	—	—	—
Six	—	—	2	—	1	—	—	—	—	—	—
Seven	—	—	—	—	1	—	—	—	—	—	—
Eight	—	—	1	—	1	—	—	—	—	—	—
Nineteen	—	—	1	—	—	—	—	—	—	—	—
<i>"Autos"</i>											
No "Autos"	472	562	288	160	56	80	49	60	101	76	213
One	77	57	43	18	20	15	22	17	6	10	76
Two	27	6	1	2	6	4	9	2	—	3	21
Three	4	3	2	1	2	1	—	1	—	—	5
Four	1	—	—	—	4	1	—	—	—	—	—
Five	—	—	—	—	2	—	—	—	—	—	—
Seven	—	—	—	—	1	—	1	—	—	—	—
<i>"Einal"</i>											
No "Einal"	486	492	277	139	61	86	57	68	76	72	271
One	89	114	52	30	33	13	22	12	30	15	41
Two	3	19	5	12	5	2	2	—	1	1	2
Three	2	3	—	—	2	—	—	—	—	—	1
Four	1	—	—	—	—	—	—	—	—	—	—

- From A.Q. Morton Paul, the man and the myth (1966)

## LLC: ce qui compte, c'est de compter

- Les objets et les résultats principes sont
- Les concordances (un objet en lui même)
- Des statistiques riches et complexes
- Il y a une hypothèse plus ou moins explicite que le “style” ou “registre” seraient identifiable de manière statistique
- Aux Etats Unis, histoire de “cliometrics” et Time on the Cross (1974)

# LLC est également une revue et un colloque

## Contents

	Volume 1 No. 1
Once. A Test of Authorship Based on Words which are not Repeated in the Sample A. G. MORTON	1
Modal Verbs and Moral Principles: An Aspect of Jane Austen's Style J. F. BURROWS	9
Test-Score Semantics as a Basis for a Computational Approach to the Representation of Meaning L. A. ZADEH	24
Text Processing in the Leningrad Research Group 'Speech Statistics' — Theory, Results, Outlook. R. G. PIOTROWSKI	36
A Computational Study of Sardinian Based upon the Proverbs Published by Canon Giovanni Spano (1871) A. GRIFFITHS	41
Programmes de Lexicométrie Syntagmatique P. LAFON, A. SALEM and M. TOURNIER	45
Diary	47
News and Notes	48
Addresses of Chairmen of Specialist Groups	50
Reviews	52
Notes for Contributors	54

Volume 5 · Number 1 · 1990

## Contents

Proper Nouns as Proper Style-Markers of Poetry and Prose M. JUILLARD	1
The Poem in Arabic, Hebrew and English and Machine Translation J. ROSENHOUSE and A. M. COHEN	9
Personal Librarian: A Tool for the Literature Classroom D. S. MALL	19
L'Informatique et les Humanités. Bibliographie 1986-1989, d'après quelques périodiques spécialisés R. PELLER and J. PRADINES	24
Special Section on Computers and Language Introduction by MARILYN DEEGAN	36
Language and IT: Rivals or Partners? F. KNOWLES	38
The Uses of Spoken and Written Corpora in the Teaching of Language and Linguistics G. KNOWLES	45
Fact—Information—Data—Knowledge: Databases as a Way of Organizing Knowledge L. HUNTER	49
The Harlib Papers Project: Text Retrieval with Large Datasets M. LESLIE	58
Categorial Grammar, Generative Phonology, and the Morphology of Old English M. DEEGAN	70
Is there a Teacher in this Class? English Language and the STELLA project at Glasgow University C. J. KAY and J. J. SMITH	77
Teaching with the Oxford Concordance Program T. T. L. DAVIDSON	81
OCP and the Computer Analysis of Texts: the Birmingham Polytechnic Experience H. JACKSON	86

<http://llc.oxfordjournals.org/>

# LLC est vivant et bien vivant en France

- Text considéré comme un phénomène statistique
- Maurice Tournier Les mots de mai 68
- Analyse factorielle et fouille de données
- Applications marketing
- Textometrie



# Humanities Computing



## 1980-1994

- Institutionalisation
- Les historiens réinvestissent le champs
- Humanities Computing : une Discipline Universitaire ?
- Projet “text encoding”

## Années 80 : décennie d'une foi illimitée dans les technologies

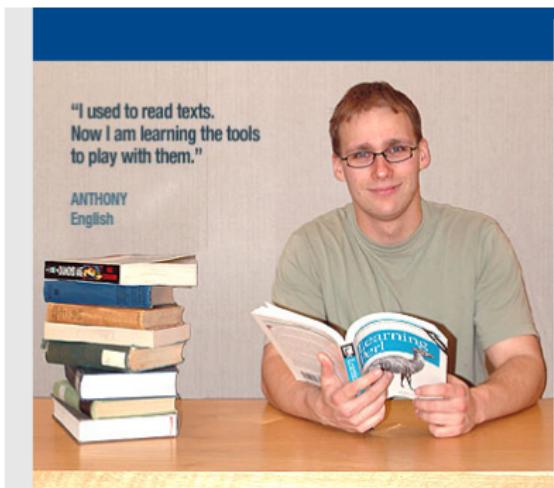
- Dans les universités les ressources et méthodes numériques bien que perçues comme étranges et difficiles trouvaient aussi leur place
- Au Royaume Uni
  - Computers in Teaching Initiative
  - Arts and Humanities Data Service
- Une nouveauté ou une amélioration du passé?
- L'arrivée du centre HC

# Communautés

- E-mail et listes de diffusion : Humanist
- Paradigmes de texte électronique
  - Oxford Text Archive, Projet Gutenberg
  - Publishing sur CD-ROM : OED
- Traitement de langage naturel et intelligence artificiel
- Financement public important pour des activités d'infrastructure
- A la fois national and européen
- Peut on gagner de l'argent dans la publication électronique?  
L'informatique personnelle? Sur l' Internet?

# Institutionalisation

- De nouveaux instances d'enseignement et de support pour l'application de l'informatique aux SHS apparaissent
- En faisant le bilan, on les valorise...



## M.A. IN **HUMANITIES COMPUTING** AT THE UNIVERSITY OF ALBERTA



The Masters of Arts in Humanities Computing is an interdisciplinary programme of the Faculty of Arts at the University of Alberta. The program integrates computational methods and theories with research and teaching in the Humanities. It addresses the demand for Arts graduates proficient in computing skills, able to work either in the realm of humanities research and teaching or in the emerging job markets of information management and content delivery over the Internet.

## Réapparition de la quellenkritik

- En France, J-P Genet et d'autres proposent l'idée que les données historiques une fois numérisées pourraient servir à enrichir une analyse
- Encore systematisée en Allemagne par Manfred Thaller avec le logiciel kleio, un sgbd textuel avant la lettre
- Une Association for History and Computing nait en 1987

## Défi pour le HC

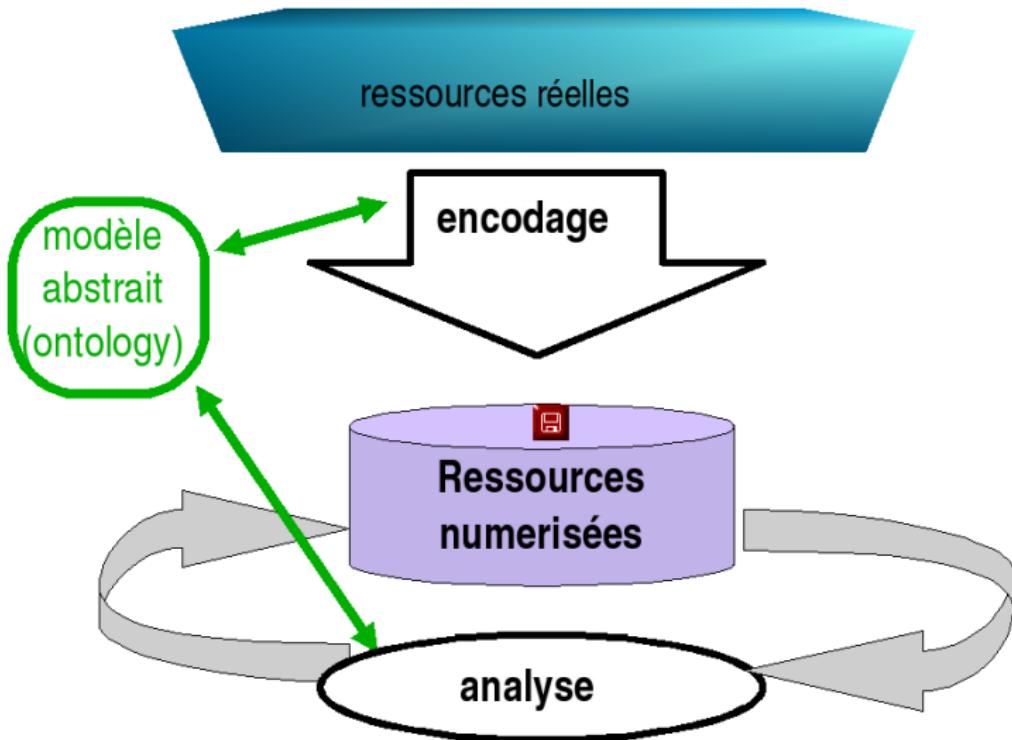
- Mais Humanities Computing ne possède aucune théorie sousjacente!
- Quel principe peut on identifier pour justifier la mise en relation des outils employés par le HC ?
- On propose les traditions scientifiques ("scholarly primitives")
  - **La recherche** effectuée selon des traits externes
  - **L'analyse** selon des traits internes
  - **Les associations** selon des perceptions partagées
- Ce qui serviraient à valoriser et contrôler l'efficacité des outils proposés

It's all about modelling, stupid



*Ceci n'est pas une pipe.*

magritte



## Les inconvenients des ressources numériques (circa 1989)

- Elles ne fonctionnent pas
- (Il faut bien choisir son ordinateur pour les faire fonctionner)
- Elles sont difficiles à trouver
- Elles ne sont pas disponibles en BU, ni mentionnées dans les revues
- Elles ne sont pas cataloguées de manière cohérente, ou pas du tout
- Elles ne sont pas fiables
- Elles émergent d'un contexte inconnu, pas (toujours) très scientifique
- Elles bougent tout le temps
- Elles ne restent pas au même endroit
- Elles disparaissent, se transforment, ou deviennent inutilisables sans préavis
- Et surtout...

Elles utilisent vraiment trop de plusieurs formats d'encodage mutuellement incompréhensibles !!

## L'effet Babel

Bien sûr il existe plusieurs lectures possibles pour la plupart des textes...

### I

#### *Loomings*

Call me Ishmael. Some years ago – never mind how long precisely – having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the

... et (malheureusement) plusieurs manières d'expression pour ces lectures!

# Encodage ou babel?

```
|chap1
<C 1> Loomings
\chapter[1]{Loomings}
:h1.1. Loomings
MOBY001001LOOMINGS
|c1
.chapter Loomings
.cp;.sp 6 a;.ce .bd 1. Loomings
<h1>Loomings</h1>
<p class="h1">Loomings
~x
```

- Bonne nouvelle: il existe des logiciels capables de traduire entre 500 formats divers
- Mauvaise nouvelle: on en a besoin

## Encodage ou babel?

```
|chap1
<C 1> Loomings
\chapter[1]{Loomings}
:h1.1. Loomings
MOBY001001LOOMINGS
|c1
.chapter Loomings
.cp;.sp 6 a;.ce .bd 1. Loomings
<h1>Loomings</h1>
<p class="h1">Loomings
~x
```

- Bonne nouvelle: il existe des logiciels capables de traduire entre 500 formats divers
- Mauvaise nouvelle: on en a besoin

## Encodage ou babel?

```
|chap1
<C 1> Loomings
\chapter[1]{Loomings}
:h1.1. Loomings
MOBY001001LOOMINGS
|c1
.chapter Loomings
.cp;.sp 6 a;.ce .bd 1. Loomings
<h1>Loomings</h1>
<p class="h1">Loomings
~x
```

- Bonne nouvelle: il existe des logiciels capables de traduire entre 500 formats divers
- Mauvaise nouvelle: on en a besoin

# La science repose sur une continuité des connaissances

- Conserver les “bytes” d'un encodage ne suffit pas
- Il faut aussi une continuité de compréhension: l'encodage doit être auto-descriptif
- Transmettre nos interprétations

D'où l'importance de laTEI (Text Encoding Initiative)

<http://www.tei-c.org>

## TEI: le résultat le plus significatif de HC?

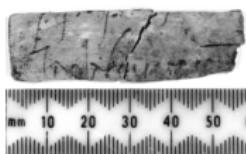
- D'origine une réponse aux problèmes posés par l'incohérence des formats et le manque des standards numériques
- La TEI est devenue un seul modèle encyclopédique des "particularités significatives" des ressources textuelles
- Et une infrastructure capable de répondre aux besoins et priorités évolutifs de la communauté scientifique

# Exemple: les tablettes Vindolanda

File Edit View Document Tools Window Help  
Side 1 x 22 / 51 88.1% Find

A Virtual Research Environment for the Study of Documents and Manuscripts

Roman fort of Vindolanda on Hadrian's Wall



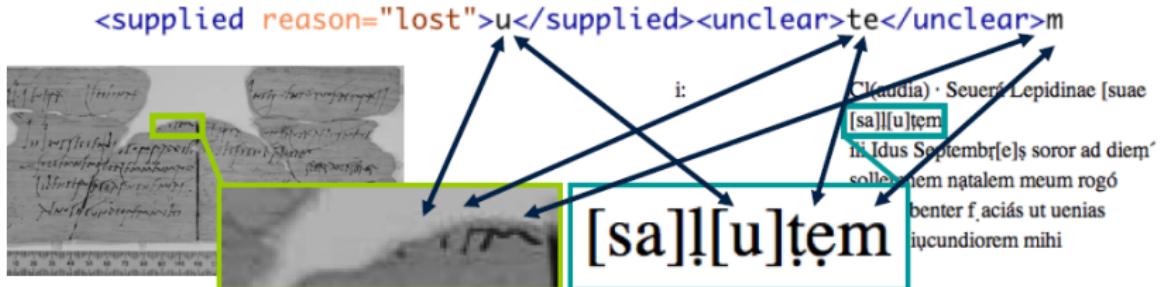
around 600 Ink and 150 Stylus Writing Tablets unearthed in 1980s

mm 10 20 30 40 50

22

# TEI en pratique

- Travaux collaboratifs de transcription scientifique
- Des conventions Leiden au standard Epidoc



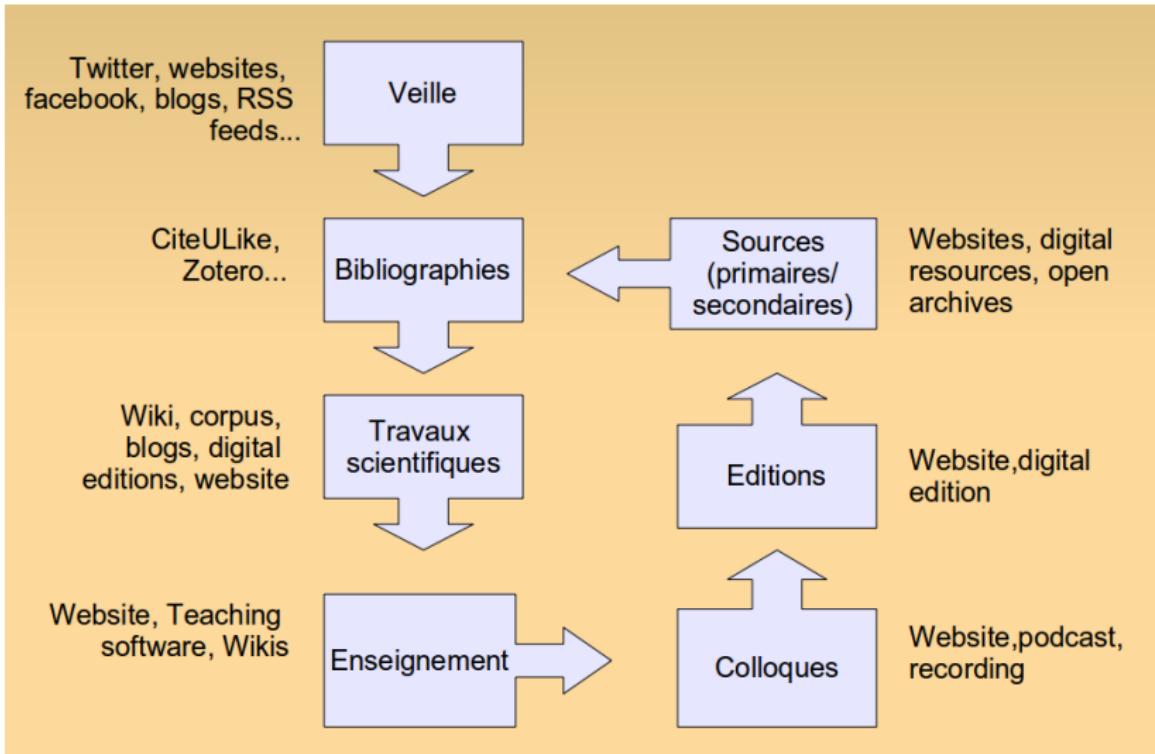
# Digital Humanities



# 1995 - ?

- Pendant que nous théorisions...
  - Le web est arrivé!
  - Le tournant numérique transforme les archives et les bibliothèques
  - La numérisation de masse s'effectue
  - Les traitements numériques se déplacent sur des grilles de services, et/ou des systèmes domestiques
  - Les réseaux sociaux emergent sur Internet
- Convergence et travaux collectifs : méthodes 'scientifique'
  - On s'interroge sur, par exemple, l'édition classique, et les méthodes collectives (cloud/crowd computing)
  - On s'aperçoit du besoin des infrastructures numériques

# La cycle de vie scientifique



## Les humanités numériques sont partout

- Comme M Jourdain, je fais des digital humanities sans le savoir?
- Les DH ne sont-elles qu'une gamme de technologies à la mode?
- "When the mode of the music changes, the walls of the palace shake"

## Digital humanities Manifesto 2.0

Digital Humanities is not a unified field but **an array of convergent practices** that explore a universe in which: a) *print is no longer the exclusive or the normative medium* in which knowledge is produced and/or disseminated; instead, print finds itself absorbed into new, multimedia configurations; and b) *digital tools, techniques, and media have altered the production and dissemination of knowledge* in the arts, human and social sciences.

<http://dev.cdh.ucla.edu/digitalhumanities/2009/05/29/the-digital-humanities-manifesto-20/#0>

## Les manifestes, ça on les connaît...

## Définition

- 1. Le tournant numérique pris par la société modifie et interroge les conditions de production et de diffusion des savoirs....
- 3. Les digital humanities désignent une transdiscipline, porteuse des méthodes, des dispositifs et des perspectives heuristiques liés au numérique dans le domaine des Sciences humaines et sociales..

## Nous constatons...

- que se sont multipliées les expérimentations dans le domaine du numérique en SHS depuis un demi-siècle ;
- que le numérique induit une présence plus forte des contraintes techniques et donc économiques dans la recherche ; que cette contrainte est une opportunité pour faire évoluer le travail collectif ;
- qu'il existe un certain nombre de méthodes éprouvées, inégalement connues et partagées ;
- qu'existent de multiples communautés particulières issues de l'intérêt pour des pratiques, des outils ou des objets transversaux divers

## Déclarations

- Nous, acteurs des digital humanities, nous nous constituons en communauté ... sans frontières. ... multilingue et multidisciplinaire.
- Nous avons pour objectifs ... l'enrichissement du savoir et du patrimoine collectif, au-delà de la seule sphère académique.
- Nous appelons à l'intégration de la culture numérique dans la définition de la culture générale du XXIe siècle.

## The economics of abundance

Digital Humanities implies the multi-purposing and multiple channeling of humanistic knowledge: no channel excludes the other. Its economy is abundance based, not one based upon scarcity ... though notions of humanistic research are everywhere under institutional pressure, there is (potentially) plenty for all. And, indeed, there is plenty to do.

## L'importance de ne pas lire

- "What can you do with a million books?" (Greg Crane)
- "Although there is still a need for close-reading... we never don't not read" (John Unsworth)
- Une nouvelle synthèse de méthodes :
  - Linguistique de corpus
  - Reconnaissances des patrons
  - Data mining
  - Visualisation
- ou une réappropriation des techniques anciens?

# Le défi: comprendre l'énormité des données disponibles

- Quelques outils disponibles aujourd'hui pour traiter un million de livres:
  - <http://books.google.com/ngrams>
  - <http://www.etalab.gouv.fr/>
  - <http://rechercheisidore.fr/>
- Mais bouger de l'étude de l'œuvre à l'étude du contexte reste problématique pour certains ...

# <http://www.scottishcorpus.ac.uk/corpus/diaview/>

Linguistic data from [Google Books Ngrams<sup>1</sup>](#). 98,098,832,753 tokens in 1851 - 2008. This is a work in progress. If you like this, then look at the [Collocate Cloud](#) ([details](#))

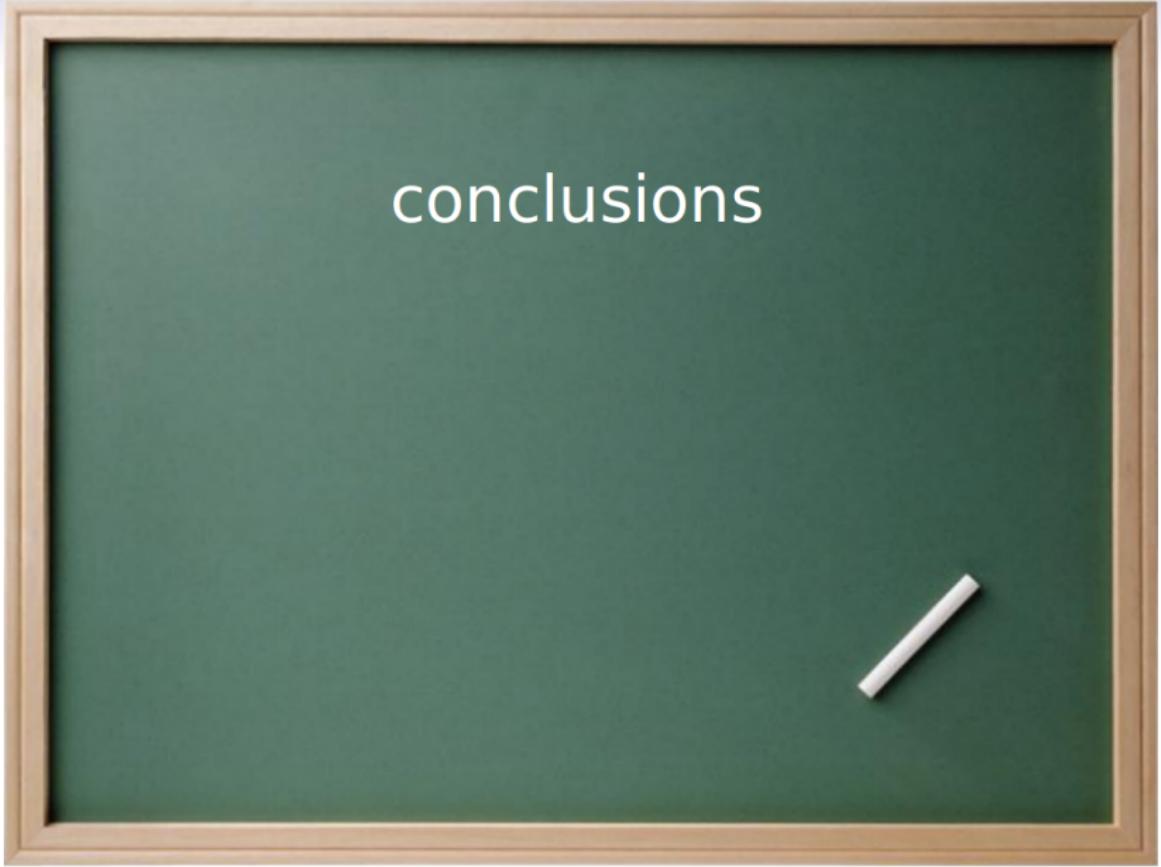
[Twitter](#) 3 [+1](#) [Email](#) David Beavan +1'd this

Show: [Overview](#) | [All years](#)

**start - end:** salient words in decreasing order (click to examine 25 years either side)

1850-1854: esa respecting saviour apostle holiness scriptures endeavoured scripture desirous countenance occasioned amidst excite alluded circumstance affective  
1855-1859: esa saviour respecting labours labors commenced aud alluded occasioned yon the affections apostle countenance desirous bosom commencement pie  
1860-1864: the esq and saviour bo rebel rebels lu commenced labours hath alluded yon doth regiments batteries slavery cavalry respecting lordship ho wellington  
1865-1869: rebel tho rebels aud yon bo gen esq saviour lu brigade ho regiment apostle commenced cavalry honorable batteries holiness alluded respecting regime  
1870-1874: honorable aud tho bo saviour lu yon signifies respecting baptism affections apostles apostle sv labour commenced esq alluded truths ut commence  
1875-1879: honorable svo aud yon lu bo esq the apostles apostle heathen cloth epistle commencement respecting saviour proprietors labours alluded crowwell co  
1880-1884: svo honorable aud lu yon mucous cloth bo inflammation crown the affections illustrations epistle esq aforesaid dd alluded heathen pulpit irritation mis  
1885-1889: svo aud plaintiff iu defendant hist tho defendants bo purchaser dd id deed crown yon honorable preached illustrations afterward heathen writ sheriff o  
1890-1894: svo aud lu honourable prof monsieur columbus yon magnet illustrations madame dd fibres lbs sensations stanley heathen crown napoleon nay marvel  
1895-1899: monsieur qu madame francs savages cried countess prof illustrations morrow qui aud fibres thither svo dante lad ye pour sulphuric nay honourable s  
1900-1904: monsieur francs madame duchess countess illustrations marquis cried morrow dante savages princess afterward tis thither mistress webster sulphur  
1905-1909: cr dante signifies cement boiler tuberculosis bee piston vine ore steam mucous lbs pipes sulphuric sulphate civilisation shaft cubic valve hydrochlor  
1910-1914: township panama bacteria tuberculosis cc sanitary sulphuric lbs hydrochloric freight railways practically cubic starch cents sulphate pipes railroad boi  
1915-1919: belgium germans cc prussian township german ideals neutrality tuberculosis austria panama prussia germany sanitary practically railroads hydrochlor  
1920-1924: cc battalion ideals practically freight instinct instincts railroads tuberculosis sulphuric lumber coal payable bonds grades detroit oils holder hydrochlor  
1925-1929: vocational pupils grades junior alpha grade superintendent pupil automobile valuation ideals cc depreciation definitely practically advertising definite  
1930-1934: definitely pupils vocational unemployment banking junior grades outstanding purchasing lease automobile alpha definite ff pupil securities shaw res  
1935-1939: definitely automobile hitler outstanding propaganda cc grid unemployment consumers retail purchasing vocational pupils league securities bureau con  
1940-1944: nazi hitler aircraft planes grid tanks radio germans vitamin propaganda aluminum voltage automobile orchestra pilot gear altitude civilian coil polish  
1945-1949: grid nazi hitler voltage aircraft orchestra radio aluminum planes roosevelt germans civilian outstanding vitamin frequencies propaganda tanks rubber  
1950-1954: communists voltage grid communist electrons communism hitler roosevelt atomic vitamin aluminum personnel orchestra outstanding nickel mi soviet  
1955-1959: communists shri atomic communist electrons communism personnel ltd eq roosevelt vitamin outstanding voltage ml inc soviet atom cit ions ussr elec  
1960-1964: communist communists eq shri communism ussr electrons ml atomic soviet electron radiation equations scattering amino amplitude ions nuclear freq  
1965-1969: shri pakistan vietnam communist nego eq communists negroes kennedy churchill rs communism electron electrons ct soviet nuclear amino ussr enz  
1970-1974: shri vietnam pollution pakistan urban behavioral blocks In enzyme variables communist programs organizational kennedy behavior plasma housing pl  
1975-1979: shri behavioral pollution In enzyme blocks variables vietnam percent singh organizational urban plasma overall oriented alternatives nuclear dna eva  
1980-1984: shri behavioral inflation nuclear plasma percent computer organizational overall processing technological variables involvement availability technolog  
1985-1989: computer software processing nuclear eds behavioral syndrome technology parameters renal strategies plasma sector implementation overall cogniti  
1990-1994: software strategies gender eds implementation options cognitive global networks dna user processing environmental parameter computer parameters  
1995-1999: gender global strategies directory eds environmental software cognitive options ethnic humans networks focused asian implementation gene uk intera  
2000-2004: web global gender options software uk user strategies phone networks users kids gene eds humans focused implementation com option files network

<sup>1</sup>Jean-Baptiste Michel\*, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven M. Stolte, and Erez Lieberman Aiden. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* (Published online ahead of print: 12/16/2010)



conclusions

# Quel est ce bruit dans la bibliothèque numérique ?

Traiter un texte c'est plus que le lire, plus que l'annoter, plus que l'associer avec d'autres textes.

Il s'agit d'exposer sa structuration afin de permettre un monde distribué, où “*les livres dans la bibliothèque savent se parler entre eux*”



## Comment effectuer une telle démarche?

- It's not rocket science (pas besoin d'avoir fait saint cyr) !
- Un balisage riche et sémantique ( par exemple TEI-XML)
- Une politique d'Open Access
- Une infrastructure permettant l'intégration et l'archivage pérenne des données

## Repenser l'édition numérique

- On est dans un monde où les documents prolifèrent, mais les textes risquent de disparaître
- Nous avons besoin de conserver nos interprétations, nos lectures, pour construire les éditions numériques
- Sans perdre les vertus traditionnelles d'un empirisme sceptique

# Composants de l'édition numérique

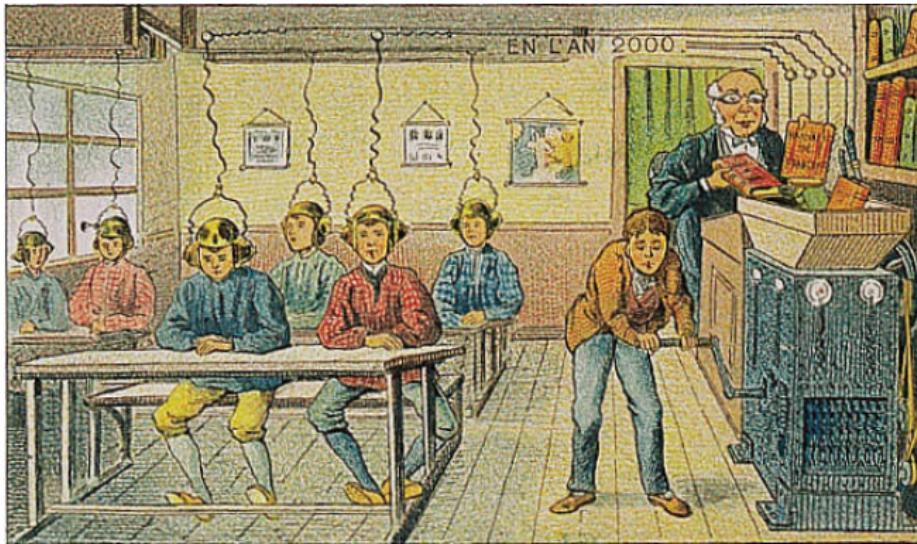
- Images de pages (ou d'autres surfaces)
- Transcriptions, éventuellement annotées
- Edition/s synthétiques
- Traduction modernes, sommaires
- Annotations paratextuelles, glossaires, préfaces, bibliographie...
- Descriptions des sources; métadonnées
- Pointeurs sur des "Factoids"

# Convergence

- Le numérique nous permet, voire oblige, d'en faire des mash up : par exemple de combiner :
  - Un SIG sur les lieux dans la mer d' Aégéan
  - Un index cartographique des toponymes de la même région
  - Un corpus de textes où ces toponymes sont attestés
  - (La TEI traite maintenant et les entités nommées et leur noms)
- De telles activités nécessitent des compétences philologiques, a priori nonautomatisable
- Et une politique d'accès ouvert

# Un rôle majeur pour les SHS

- Nous comprenons les objets textuels
  - De quelle manière se présente ce discours?
  - Quelles sont les histoires qu'il raconte?
- Nous connaissons l'hermeneutique
  - quelle est la portée de ce discours?
  - Qu'est-ce qu'il veut dire – mais ne dit pas ?
- Voici notre contribution au web sémantique.



Merci de votre attention!

